Cuestionario de teoría 3

Antonio Álvarez Caballero
analca3@correo.ugr.es

12 de junio de 2016

1. Cuestiones

Cuestión 1. Considera los conjuntos de hipótesis $\mathcal{H}_1$ y $\mathcal{H}_{100}$, que contienen funciones Booleanas sobre 10 variables Booleanas, es decir $\mathcal{X} = \{-1, +1\}^{10}$. $\mathcal{H}_1$ contiene todas las funciones Booleanas que toman valor +1 en un único punto de $\mathcal{X}$ y -1 en el resto. $\mathcal{H}_{100}$ contiene todas las funciones Booleanas que toman valor +1 exactamente en 100 puntos de $\mathcal{X}$ y -1 en el resto.

a) ¿Cuántas hipótesis contienen $\mathcal{H}_1$ y $\mathcal{H}_{100}$?

b) ¿Cuántos bits son necesarios para especificar una de las hipótesis en $\mathcal{H}_1$?

c) ¿Cuántos bits son necesarios para especificar una de las hipótesis en $\mathcal{H}_{100}$?

Argumente sobre la relación entre la complejidad de una clase de funciones y la complejidad de sus componentes.

Solución. Respondamos a cada cuestión:

a) $\mathcal{H}_1$ contiene 2^{10} hipótesis. Esto es porque para una función booleana con 10 entradas, obtenemos $2^{10} = 1024$ salidas. Como queremos 1 en una entrada y -1 en lo demás, este número coincide con el número de posibles entradas. Para el caso $\mathcal{H}_{100}$ razonamos igual, aunque esta vez debemos coger de las 1024 posibilidades, subconjuntos de 100 1. Por tanto, este es el número combinatorio $\binom{1024}{100} = 7{,}7466\ldots \times 10^{140}$.

b) Para especificar una de las hipótesis de $\mathcal{H}_1$ sólo nos hace falta especificar el punto donde vale +1, luego con 10 bits es suficiente.

c) Al igual que antes, necesitamos especificar dónde vale +1 la hipótesis. Como tenemos 100 +1, necesitamos $100 \cdot 10 = 1000$ bits.

La complejidad de una clase de funciones está directamente relacionada con la de sus componentes. A más complejidad en las componentes, más compleja será la clase de funciones. Le estamos dando más grados de libertad a las funciones del conjunto de hipótesis cuanto más aumentamos la complejidad de las componentes.

Cuestión 2. Suponga que durante 5 semanas seguidas, recibe un correo postal que predice el resultado del partido de fútbol del domingo, donde hay apuestas sustanciosas. Cada lunes revisa la predicción y observa que la predicción es correcta en todas las ocasiones. El día de después del quinto partido recibe una carta diciéndole que si desea conocer la predicción de la semana que viene debe pagar 50.000€. ¿Pagaría?

a) ¿Cuántas son las posibles predicciones gana-pierde para los cinco partidos?

b) Si el remitente desea estar seguro de que al menos una persona recibe de él la precicción correcta sobre los 5 partidos, ¿Cuál es el mínimo número de cartas que deberá de enviar?

c) Después de la primera carta prediciendo el resultado del primer partido, ¿A cuántos de los seleccionados inicialmente deberá enviarle la segunda carta?

d) ¿Cuántas cartas en total se habrán enviado después de las primeras cinco semanas?

e) Si el coste de imprimir y enviar las cartas es de 0.5 por carta, ¿Cuánto ingresa el remitente si el receptor de las 5 predicciones acertadas decide pagar los 50.000€?

f) ¿Puede relacionar esta situación con la función de crecimiento y la credibilidad del ajuste de los datos?

Solución. No pagaría. Veamos poco a poco por qué.

a) Suponiendo que los partidos no pueden empatar, las posibles predicciones gana-pierde son claramente $2^5 = 32$ posibilidades.

b) Al ser un problema binario, el número mínimo de cartas a enviar es $\sum_{i=1}^{5} 2^i$. Esto es porque la primera semana se deben enviar $2^5 = 32$ cartas, la mitad de ellas con una posibilidad (gana A) y la otra mitad con la otra (gana B). La siguiente semana sólo la mitad de ellos tendrán la predicción correcta, luego hay que enviar $2^4 = 16$ cartas a los que obtuvieron la predicción correcta, igualmente con mitad y mitad. Así hasta la 5º semana, que sólo habrá que enviar $2^1 = 2$ cartas, asegurando a una persona haber recibido 5 predicciones correctas.

c) Como se ha explicado antes, hay que enviarle cartas a las personas que hayan recibido la predicción correcta. En el caso anterior, para asegurarnos que una persona reciba después de 5 semanas las predicciones correctas, hay que enviar cartas a la mitad de personas, $2^4 = 16$.

d) En total se mandan $\sum_{i=1}^{5} 2^i = 62$. Si contamos la última carta, la que le manda al receptor de las 5 predicciones correctas, pidiendo los 50.000€, en total se mandan 63 cartas.

e) Se habría gastado 31€ y habría ganado 50.000€, luego el beneficio sería de 49.969€. Contando la última carta, serían 50 céntimos menos de beneficio.

f) El espacio de hipótesis es desconocido para todos los receptores de cartas. Para estas personas, mientras vaya acertando el predictor, la función de crecimiento es 1, por lo la credibilidad en el ajuste es total. Pero en realidad, la función de crecimiento de este modelo es 2^N, siendo N el número de partidos a predecir, por lo que la credibilidad del ajuste es muy baja al ser baja la probabilidad de que se dé el resultado esperado.

Cuestión 3. En un experimento para determinar la distribución del tamaño de los peces en un lago, se decide echar una red para capturar una muestra representativa. Así se hace y se obtiene una muestra suficientemente grande de la que se pueden obtener conclusiones estadísticas sobre los peces del lago. Se obtiene la distribución de peces por tamaño y se entregan las conclusiones. Discuta si las conclusiones obtenidas servirán para el objetivo que se persigue e identifique si hay algo que lo impida.

Solución. El objetivo no tiene por qué conseguirse, porque a priori no sabemos si la muestra es realmente representativa. Si la distribución de tamaño de los peces del lago no coincide con la distribución que sigue la muestra, el objetivo no se cumplirá, porque hemos estado tomando una muestra sesgada de la población de peces. Factores como la época del año, el tamaño y forma de la red o el lugar donde se realiza la prueba son idóneos para sesgar la muestra.

Cuestión 4. Considere la siguiente aproximación al aprendizaje. Mirando los datos, parece que los datos son linealmente separables, por tanto decidimos usar un simple perceptrón y obtenemos un error de entrenamiento cero con los pesos óptimos encontrados. Ahora deseamos obtener algunas conclusiones sobre generalización, por tanto miramos el valor d_{VC} de nuestro modelo y vemos que es $d + 1$. Usamos dicho valor de d_{VC} para obtener una cota del error de test. Argumente a favor o en contra de esta forma de proceder identificando los posibles fallos si los hubiera y en su caso cuál hubiera sido la forma correcta de actuación.

Solución. El principal problema es que hemos visto los datos. Al haber visto los datos ya tenemos información del problema, lo cual sesga el conocimiento que podemos aprender de ellos. Así se pierde la capacidad de generalización del modelo.

El tema de la cota pierde totalmente su validez al haber contaminado el aprendizaje, luego esa cota tampoco será real.

Una forma correcta de actuar sería tomar un modelo, analizar sus errores dentro de la muestra y de test y ya con la d_{VC} calcularíamos una cota del error de generalización. Esto aplicado a varios modelos podrá darnos una estimación más fiable del modelo más apropiado para este problema.

Cuestión 5. Suponga que separamos 100 ejemplos de un conjunto $\mathcal{D}$ que no serán usados para entrenamiento, sino que serán usados para seleccionar una de las tres hipótesis finales g_1, g_2, g_3 producidas por tres algoritmos de aprendizaje distintos entrenados sobre el resto de datos. Cada algoritmo trabaja con un conjunto $\mathcal{H}$ de tamaño 500. Nuestro deseo es caracterizar la precisión de la estimación $E_{out}(g)$ sobre la hipótesis final seleccionada cuando usamos los mismos 100 ejemplos para hacer la estimación.

a) ¿Qué expresión usaría para calcular la precisión? Justifique la decisión.

b) ¿Cuál es el nivel de contaminación de estos 100 ejemplos comparándolo con el caso donde estas muestras fueran usadas en el entrenamiento en lugar de en la selección final?

Solución. Resolvemos por partes.

a) Como el conjunto $\mathcal{H}$ es finito, la expresión que debemos utilizar para estimar E_{out} es la desigualdad de Hoeffding. Lo único que tenemos que ajustar es $|\mathcal{H}|$. Es claro que $|\mathcal{H}| = \sum_{i=1}^{3} |\mathcal{H}_i| = 1500$, siendo $\mathcal{H}_i$ el espacio de hipótesis para cada g_i. No es sólo 3, ya que las tres funciones han sido propuestas por los propios algoritmos de aprendizaje, por lo que previamente han tenido que ser escogidas de cada uno de los espacios $\mathcal{H}_i$, que son de tamaño 500.

b) La contaminación es mayor si elegimos la función hipótesis en base a estos 100 ejemplos, ya que dependerá en mayor medida de estos 100 ejemplos. Si hubiéramos dejado dentro estos 100 elementos, habrán tenido peso en la elección de una hipótesis, o incluso contaminaríamos el aprendizaje como en el ejercicio anterior (el modelo conoce los datos), pero posiblemente esta contaminación sea más indirecta.

Cuestión 6. Considere la tarea de seleccionar una regla del vecino más cercano. ¿Qué hay de erróneo en la siguiente lógica que se aplica a la selección de k? (Los límites son cuando $N \to \infty$). Considere la posibilidad de establecer la clase de hipótesis $\mathcal{H}_{NN}$ con N reglas, las k-NN hipótesis, usando $k = 1, \ldots, N$. Use el error dentro de la muestra para elegir un valor de k que minimiza E_{in}. Utilizando el error de generalización para N hipótesis, obtenemos la conclusión de que $E_{in} \to E_{out}$ porque $\frac{\log(N)}{N} \to 0$. Por lo tanto concluimos que asintóticamente, estaremos eligiendo el mejor valor de k, basándonos sólo en E_{in}.

Solución. Es claro que el k elegido sería $k = 1$, ya que para dicho k y tomando los datos dentro de la muestra, $E_{in} = 0$. Sabemos que $k = 1$ no es el mejor valor para el k-NN, por lo que esta lógica para elegir k no es la adecuada. Sería más adecuado tomar $k = \sqrt{N}$ o bien utilizar validación cruzada.

Cuestión 7. Responder estas cuestiones:

a) Considere un núcleo Gaussiano en un modelo de base radial. ¿Qué representa $g(x)$ (ecuación 6.2 del libro LfD) cuando $||x|| \to \infty$ para el modelo RBF no-paramétrico vs el modelo RBF paramétrico, asumiendo w_n fijos?

b) Sea $\mathcal{Z}$ una matriz cuadrada de características definida por $\mathcal{Z}_{nj} = \Phi_j(x_n)$, donde $\Phi_j(x)$ representa una transformación no lineal. Suponer que $\mathcal{Z}$ es invertible. Mostrar que un modelo paramétrico de base radial, con $g(x) = w^T\Phi(x)$ y $w = \mathcal{Z}^{-1}y$, interpola los puntos de forma exacta. Es decir, que $g(x_n) = y_n$, con $E_{in}(g) = 0$.

c) ¿Se verifica siempre que $E_{in}(g) = 0$ en el modelo no paramétrico?

Solución. solucion

Cuestión 8. Verificar que la función *sign* puede ser aproximada por la función *tanh*. Dado w_1 y $\epsilon > 0$ encontrar w_2 tal que $|sign(x_n^T w_1) - tanh(x_n^T w_2)| \leq \epsilon$ para $x_n \in \mathcal{D}$. Ayuda: Analizar la función $tanh(\alpha x), \alpha \in \mathbb{R}$.

Solución. Hagamos en primer lugar un pequeño análisis de la función $tanh(\alpha x), \alpha \in \mathbb{R}$. Veamos qué pasa cuando $\alpha \to \infty$. Por la forma de la tangente hiperbólica, podemos esperar qué va a pasar.

$$\lim_{\alpha \to +\infty} tanh(\alpha x) = \lim_{\alpha \to +\infty} \frac{e^{\alpha x} - e^{-\alpha x}}{e^{\alpha x} + e^{-\alpha x}} = 1$$

$$\lim_{\alpha \to -\infty} tanh(\alpha x) = \lim_{\alpha \to -\infty} \frac{e^{\alpha x} - e^{-\alpha x}}{e^{\alpha x} + e^{-\alpha x}} = -1$$

Deducimos entonces que $sign(x) \approx tanh(\alpha x),\ \alpha \to \infty$. Vamos a tomar $w_2 = \alpha \cdot w_1$. Podemos hacer dicha diferencia tan pequeña como queramos (ϵ) aumentando α todo lo que queramos hasta alcanzar dicho ϵ.

Cuestión 9. Sea V y Q el número de nodos y pesos en una red neuronal,

$$V = \sum_{l=0}^{L} d^{(l)}, \quad Q = \sum_{l=1}^{L} d^{(l)} \left(d^{(l+1)} + 1 \right)$$

En términos de V y Q, ¿Cuántas operaciones se realizan en un pase hacia adelante (sumas, multiplicaciones, y evaluaciones de θ)? Ayuda: Analizar la complejidad en términos de V y de Q.

Solución. Como para una multiplicación matriz-vector hay que realizar una multiplicación por cada elemento de la matriz, entonces hay una multiplicación por cada uno de los pesos de la red. Luego hay Q multiplicaciones. En términos de complejidad es $O(Q)$.

Para las sumas en una multiplicación matriz-vector de tamaño $n \times m$ hay que realizar $n \cdot (m-1)$ sumas. Por lo que, si para cada capa hay que realizar $\left(d^{(\ell)}\right) \cdot \left(d^{(\ell-1)}\right)$ sumas, en total son $\sum_{\ell=1}^{L} d^{(\ell)} d^{(\ell+1)}$. En términos de Q, separando la sumatoria de Q, nos salen $Q - V + d^{(0)}$. En términos de complejidad es $O(Q)$, ya que Q tiene un orden de magnitud (según d) mayor que V.

Por último, las θ − evaluaciones. La función se evalúa en $\sum_{\ell=1}^{L} d^{(\ell)} = V - d^{(0)}$ ocasiones, luego en términos de complejidad es $O(V)$.

En total, el número de operaciones totales es $2Q$.

Cuestión 10. Para el perceptrón sigmoidal $h(x) = tanh(x^T w)$, sea el error de ajuste $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}\Big(tanh(x_n^T w) - y_n\Big)^2$. Mostrar que

$$\nabla E_{in}(w) = \frac{2}{N}\sum_{n=1}^{N}\Big(tanh(x_n^T w) - y_n\Big)\Big(1 - tanh(x_n^T w)^2\Big)\, x_n$$

Si $w \to \infty$, ¿Qué le sucede al gradiente? ¿Cómo se relaciona esto con la dificultad de optimizar el perceptrón multicapa?

Solución. Derivamos con respecto a w, que simplemente es aplicar la regla de la cadena varias veces:

$$\nabla E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} 2\Big(tanh(x_n^T w) - y_n\Big)\Big(1 - tanh(x_n^T w)^2\Big)\, x_n$$

Si $w \to \infty$, queda, utilizando los límites del ejercicio 8:

$$\lim_{w\to\infty} \nabla E_{in}(w) = \lim_{w\to\infty} \frac{1}{N}\sum_{n=1}^{N} 2\underbrace{\Big(tanh(x_n^T w) - y_n\Big)}_{\in\{-2,-1,0,+1,+2\}}\underbrace{\Big(1 - tanh(x_n^T w)^2\Big)}_{0}\, x_n = 0.$$

Al anularse el gradiente con $w \to \infty$, se terminará estancando en un óptimo local. Una vez el gradiente es 0, deja de iterar el algoritmo, luego para evitar eso, siempre se suele comenzar con un w muy pequeño.